REMARKS

Claims 1-5 and 7-25 are pending. Claim 6 has been canceled. Support for the newly added claims is found derives from the specification and claims as originally filed. For example, computational methods for the generation of primary libraries are described at page 6, line 31, to page 10, line 12, and page 11, line 1 to page 13, line 34. Methods for the generation of secondary libraries from primary libraries are described at page 24, line 9, to page 27, line 32. Support for the synthesis of variant proteins, beginning with the corresponding oligonucleotide sequences using multiple PCR can be found at pages 28-30 Methods for isolating, purifying, and expressing the oligonucleotide sequences as proteins are well known in the art, and are described at pages 37-44 and in the Examples. Accordingly, the amendments do not present new matter and entry is proper.

Lack of Utility:

Claims 1-6 are rejected under 35 U.S.C. §101. The Office Action asserts that no specific or well-established utility has been disclosed. Reconsideration under 37 CFR 1.111 is requested.

Applicants are claiming a **method** of generating a secondary library. The instant claims are not directed to a library or a composition of matter.

The claims of the present invention provide a method for computationally screening variant protein sequence libraries to generate secondary libraries of useful variant protein sequences, which when synthesized find use in a wide variety of applications, ranging from industrial to pharmacological uses. Furthermore, the methodology of the present invention allows for the rapid screening of large numbers of potential variant sequences for useful variants and the selection of proteins with useful properties. Greater diversity of protein sequences may be obtained by the method of the present invention. See Specification at page 2, lines 13-19; page 4, lines 31-36; page 6, lines 3-8; and page 6, lines 9-17.

A secondary library is defined as follows in the specification:

> "In a preferred embodiment, the primary library of the scaffold protein is used to generate a secondary library. As will be appreciated by those skilled in the art, the secondary library can be either a subset of the primary library, or contain new library member, i.e. sequences that are not found in the primary library. That is, in general, the variant positions and/or amino acid residues in the variant positions can be recombined in any number of ways to form a new library that exploits the sequence variations found in the primary library. That is, having "hot spots" or important variant positions and/or residues, these positions can be recombined in novel ways to generate novel sequences to form a secondary library. Thus in a preferred embodiment, the secondary library comprises at least one member sequence that is not found in the primary library, and preferably a plurality of such sequences."

The Applicants respectfully draw the Examiner's attention to the Utility Guidelines:

In most cases, an applicant's assertion of utility creates a presumption of utility that will be sufficient to satisfy the utility requirement of 35 U.S.C. § 101. As the CCPA stated in In re Langer:

> "As a matter of Patent Office practice, a specification which contains a disclosure of utility which corresponds in scope to the subject matter sought to be patented must be taken as sufficient to satisfy the utility requirement of § 101 for the entire claimed subject matter unless there is a reason for one skilled in the art to question the objective truth of the statement of utility or its scope."

Thus, Langer and subsequent cases direct the Patent Office to presume that a statement of utility made by an applicant is true. For obvious reasons of efficiency and in deference to an applicant's understanding of his or her invention, when a statement of utility is evaluated, Patent Office personnel should not begin an inquiry by questioning the truth of the statement of utility. Instead, any inquiry must start by asking if there is any reason to question the truth of the statement of utility. This can be done by evaluating the logic of the statements made, taking into consideration any evidence cited by the applicant. If the asserted utility is credible (i.e., believable based on the record or the nature of the invention), a rejection based on "lack of utility" is not appropriate. Thus, Patent Office personnel should not begin an evaluation of utility by assuming that an asserted utility is likely to be false, based on the technical field of the invention or for other general reasons.

Compliance with § 101 is a question of fact. Thus, to overcome the presumption of truth that an assertion of utility by the applicant enjoys, Patent Office personnel must establish that it is more likely than not that one of ordinary skill in the art would doubt (i.e., "question") the truth of the statement of utility. To do this, Patent Office personnel must provide evidence sufficient to show that a person of ordinary skill in the art would consider the statement of asserted utility "false". A person of ordinary skill must have the benefit of both facts and reasoning in order to assess the truth of a statement. This means that if the applicant has presented facts that support the reasoning used in asserting a utility, Patent Office personnel must present countervailing facts and reasoning sufficient to establish that a person of ordinary skill would not believe the applicant's assertion of utility (MPEP §2107.02IIIA). The initial evidentiary standard used during evaluation of this question is a preponderance of the evidence (i.e., the totality of facts and reasoning suggest that it is more likely than not that the statement of the applicant is false). It is respectfully submitted that the Examiner has not met this burden.

Applicant respectfully submits that the application is enabled by the Examples where a molecule

whose coordinates were input into a computer, heavy side chain atoms were selected within a 4 Angstrom sphere around four catalytic residues. In addition, another example shows a set of residues within a 5 Angstrom sphere are floated. A probability table (Table 3) was calculated from the top 1000 sequences in the list (again see Table 3). Table 3 shows the number of occurrences of each of the amino acids selected for each position (i.e., 5 variable positions and 25 floated positions). Thus these examples also show utility.

Additionally, no further characterization of the present invention is necessary to demonstrate or confirm a "real world" use because methods of protein design related to those of the present invention have been shown to work as claimed. See also U.S. Patent Nos. 6,188,965; 6,296,312; 6,403,312; 6,708,120; 6,792,356; PCT/US98/07254 and PCT/US01/40091. Such methods have been used to generate novel proteins with enhanced properties, see for example, U.S. Patent Nos. 6,682,923; 6,627,186; 6,514,729; and 6,746,853. See also, Steed et al, *Science* (2003), 301: 1895-1898, a copy of which is enclosed as Exhibit A; Hayes et al., *PNAS*, 99 (25): 15926-15931, a copy of which is enclosed as Exhibit B; and Luo et al., *Protein Science* (2002), 11: 1218-1226, a copy of which is enclosed as Exhibit C. Applicant also notes that the methodology described in these patents and scientific publications is not limited to enzymes, but applies to therapeutic proteins as well as any other type of protein.

In the article "Proteins from Scratch" (DeGrado, *Science* (1997), 278:80-81, a copy of which is enclosed as Exhibit D), biochemistry professor William F. DeGrado of the University of Pennsylvania School of Medicine, a world-renowned expert in protein structure, folding and design, comments on the computational platform designed by Dahiyat and Mayo in Science (1997), 278:82-87. This platform is an earlier version of the computational platform that has evolved and is claimed herein. Dr. DeGrado states:

> "Not long ago, it seemed inconceivable that proteins could be designed from scratch. Because each protein sequence has an astronomical number of potential confirmations, it appears that only an experimentalist with the evolutionary life span of Mother Nature could design a sequence capable of folding into a single, well-defined three dimensional structure. But now on page 82 of this issue, Dahiyat and Mayo describe a new approach that makes de novo protein design as easy as running a computer."

Dr. DeGrado further states (col 1, paragraph 3):

> "Thus, the problem of de novo protein design reduced to two steps: selecting a desired tertiary structure and finding a sequence that would stabilize this fold. Dahiyat and Mayo have now mastered the second step with spectacular success. They have distilled the rules, insights and paradigms gleaned from two decades of experiments into a single computational algorithm...Thus the rules of ...computational methods for de novo design may now be sufficiently defined to allow the engineering of a variety of proteins."

Further, in 2002, Dr. Jeffery G. Saven, a well-known expert in protein design, has recently published a review of the state of the art in combinatorial protein libraries (see, Saven, JG, Curr. Op. Struct. Biol. (2002), 12:453-458, a copy of which is enclosed as Exhibit E, where he states at page 456, col. 1, 3rd paragraph, lines 6 – 13:

> "Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but with combinatorial experiments, an enormous diversity of sequences can be generated at the control of the researcher".

The Saven publication, while not prior art in the instant application, shows that it is known in the art that combinatorial library generation has "real world use". Thus, the discussions above regarding examples of actual utility by Applicant, as well as recognition to those skilled in the art of protein design and combinatorial library generation, meets the utility requirement under 35 USC § 101.

As further outlined in the Guidelines:

> Where an applicant has specifically asserted that an invention has a particular utility, that assertion cannot simply be dismissed by Office personnel as being "wrong," even when there may be reason to believe that the assertion is not entirely accurate. Rather, Office personnel must determine if the assertion of utility is credible (i.e., whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided). An assertion is credible unless (a) the logic underlying the assertion is seriously flawed, or (b) the facts upon which the assertion is based are inconsistent with the logic underlying the assertion. Credibility as used in this context refers to the reliability of the statement based on the logic and facts that are offered by the applicant to support the assertion of utility.

Thus, the burden is shifted to the Examiner. The Examiner analogizes a library to a composition of matter, which has to undergo screening to isolate and identify a product, citing Brenner v. Manson, 148 USPQ 689 (1966) ("Brenner").

Applicants are specifically claiming a **method** of generating a secondary library, not a "library" per se, nor a composition of matter in the instant application. Thus, the analogy to Brenner that the Examiner makes is not analogous to the claims in the instant application.

In addition, Applicants respectfully disagree with the analogy to Brenner because the protein variants to be screened by the method of the present invention, synthesized and/or tested find utility in their respective fields. For example, for purposes of the present invention, it does not matter what the

class of proteins are. The method of the claimed invention, screens for useful variants having desired protein characteristics. See for example, Specification at page 4, lines 25-30 and page 34, lines 22 to page 35, line 12. For example, the variants produced from the method of the present invention may find use as therapeutic proteins. See Specification beginning at page 34, lines 22, ending on page 35, line 12.

The arguments made above with respect to 35 USC §101 are equally applicable to the rejection under 35 USC §112, first paragraph. The techniques described in the recited methods have a specific and well-established utility, and one skilled in the art would know how to use the claimed invention, particularly as demonstrated in the patents and scientific articles discussed above.

Thus, the burden is shifted to the Examiner. The Examiner analogizes a library to a composition of matter, which has to undergo screening to isolate and identify a product, citing Brenner v. Manson, 148 USPQ 689 (1966). Applicants respectfully disagree because the invention as claimed is in fact a method for generating libraries. Although the examiner describes the secondary libraries as presently undefined, the method for generating them is fully enabled by the specification. The basis for this is that Applicant's are claiming a method of generating a secondary library. The library generated will necessarily vary with the particular target protein identified, as well as the use of the different parameters of the method.

In conclusion, as outlined above, Applicant's are claiming a method of generating a secondary library, not a library per se, nor a composition of matter. One skilled in the art would be able to practice the invention as described in the method claims and by a review of the enabling specification. It is submitted that the present invention has utility under §101 and §112, first paragraph and Applicants respectfully request that the rejections be withdrawn.

Claim Rejection 35 USC §112, first paragraph

Claims 1-3 are rejected under 35 USC §112, first paragraph because the specification while enabling for the enzymes protein design using specific program design, does not reasonably provide enablement for any type of secondary library of scaffold protein variants or sequences.

By way of clarifying the terms used in the specification, Office Action and the instant response, Applicants would like to provide definitions of the following terms of art (as shown in Voet & Voet, Biochemistry, Chapter 6, page 109 (1990); Exhibit F:)

1. A protein's **primary structure (1° structure)** is the amino acid sequence of its polypeptide chain(s).

2. **Secondary (2°) structure** is the local spatial arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains.

3. **Tertiary (3°) structure** refers to the three-dimensional structure of an entire polypeptide. The distinction between secondary and tertiary structures is, of necessity, somewhat vague; in practice, the term secondary structure alludes to easily characterized structural entities such as helices.

The Applicants respectfully disagree for the following reasons. §112 does not require such extensive disclosure. A patent need not teach, and preferably omits, what is well known in the art. *In re Buchner*, 929F.2d 660, 661, 18 USPQ2d 1331, 1332 (Fed. Cir. 1991); *Hybritech Inc. v. Monoclonal Antibodies, Inc.*, 802 F.2d 1367, 1384, 231 USPQ81, 94 (Fed.Cir. 1986), cert.denied, 480 U.S. 947 (1987); and *Lindemann Maschinenfabrik GMBH v. American Hoist & Derrick Co.*, 730 F.2d 1452, 1463, 221 USPQ481, 489 (Fed. Cir. 1984).

Furthermore, "[a]ll that is necessary is that one skilled in the art be able to practice the claimed invention, given the level of knowledge and skill in the art. Further, the scope of enablement must only bear a "reasonable correlation" to the scope of the claims. Se, e.g., *In re Fisher*, 427 F.2d 833, 839, 166 USPQ 18, 24 (CCPA 1970)." (See MPEP §2164.08)

The Applicant respectfully draws the Examiner's attention to page 7, line 9 to page 9, line 5 of the Specification as filed, where there is a discussion of secondary library of scaffold protein or variant sequences.

The enablement requirement refers to the requirement of 35 USC 112, first paragraph that the specification describe how to make and how to use the invention. The invention that one skilled in the art must be enabled to make and use is that defined by the claim(s) of the particular application or patent.

As stated *infra* regarding utility, Applicant respectfully submits that the application is enabled by the examples where a molecule whose coordinates were input into a computer, heavy side chain atoms were selected within a 4 Angstrom sphere around four catalytic residues. These heavy side chain atoms defined the variable residue positions for which a primary library was calculated. A probability table (Table 3) was calculated from the top 1000 sequences in the list (again see Table 3). Table 3 shows the number of occurrences of each of the amino acids selected for each position (i.e., 5 variable positions and 25 floated positions). One skilled in the art would readily be capable of extrapolating these examples to a variety of protein systems with a variety of functions, particularly when read in light of the specification (e.g. see Specification page 7, line 27 to page 9, line 5; page 34, line 22 to page 35, line 12). Thus these

examples also show enablement.

With respect to the scope of the enabling disclosure not commensurate with the scope provided in the Specification, there is disclosure of using a computational design program, and preferably PDA® technology as embodiments of the invention. See Specification at page 2, lines 1-3; page 7, lines 9-12; and page 14, line 30 to page 15, line 5. In addition, the examples provide further enabling disclosure to one skilled in the art to practice this invention. As stated previously, the methodology is not limited to a particular kind of protein, and one skilled in the art would not be led to believe that this method is limited to enzymes. The method of the present invention is not limited to enzymes, since the modifications may be done to any proteins, not just enzymes. The methodology has been successfully employed in many non-enzyme proteins, e.g., TNF, GCSF, Interferon, etc. The publications cited in the section addressing the 35 USC §101 show the diversity of proteins that may be used. In addition, the article by Dr. Saven shows that those skilled in the art do not limit proteins by type (such as enzyme). The methodologies apply to any type of protein. The methodology requires that coordinates of a target protein be input. There is nothing in the methodology that so limits it only to enzymes, and while the examples show enzyme modifications, these examples are just that, examples of how the technology works. The specification provides support for the use of any protein that may be used in this method. One skilled in the art would understand that this method may be used on any protein and not just limited to enzymes.

Examiner states that "in an unpredictable art such as protein" (page 5, 2nd paragraph), that it is difficult to design a protein. Applicants are not designing a protein de novo, but are inputting the coordinates of a target protein. Inputting the coordinates of a target protein is the equivalent to enabling the analysis of that particular protein structure. Thus, Applicants are not designing a secondary or tertiary structure (see definitions of these terms *supra*) of a protein de novo, but relying on the existing structure coordinates from which to commence a design. The methodology employs known physico-chemical parameters of proteins, amino acids and rotamers to modify the target protein.

The Examiner notes that "hydrogen bonding could be important in designing and stabilizing of other types of secondary structure" (Office Action, page 6, lines 10-13) as an example of the complexity of the protein design art. Hydrogen bonding is one of the physico-chemical properties that may be used in Applicants' methodology (see Specification at page 14, line 14; page 16, line 18; page 18 line 9 to page 19, line 3; page 58, line 7). Further, hydrogen bonding is one of the scoring function types that is combined in a balanced manner to accurately model proteins (see for example p. 18, Equation 1, or US6188965 or references incorporated therein). Many scoring functions can be used successfully (see Specification page 10, line 17 to page 15 line 5) but each balances similar energy terms, such as hydrogen

bonding and secondary structure propensity, slightly differently.

The Examiner also asks "if such design is feasible in the actual environment where the protein exists" (Office Action, page 7, lines 1-2). The answer to that question is Yes. Some of the parameters of the methodology are designed to take the environment of the protein into account (e.g. see discussions of solvation energies in the Specification at page 14, line 14 and page 18, line 5). Further, more specifically, the third paragraph of Example 3 and Table 3 recite the suite of parameters that may be used to design a modified protein in its environment and in a quantitative manner.

Applicants specifically prepare a structure for design by eliminating everything but the protein backbone and eliminating "non-protein" elements (e.g. water). (see Specification at page 15, line 4 to page 16, line 8).

It is by this technique that the coordinates are input by defining the backbone of the protein. The target protein is thus prepared to for its coordinates to be input (see specification at page 7, lines 27-31 and page 15, line10 to page 16, line 8). The target protein may include any type protein (see specification at page 8, line 8 to page 9, line 5 for disclosure on non-enzyme protein scaffolds).

Next amino acid groups are selected and positions are selected at page 16, lines 9-17 in the specification. It is noted that Examiner's comments regarding the non-use of certain amino acids is addressed in detail in the specification as cited. Scoring functions are described at page 10, line 18 to page 20, line 19. Scoring functions are applied to the target protein and the selected amino acids at the selected positions to generate a library of variants that have been optimized or generated (see page 20, line 20 to page 21, line 15) to achieve a particular design goal.

Thus for every protein (not just enzymes), the same methodology as recited in the instant claims is used. The coordinates are input; a scoring function is used to generate a primary library; a probability distribution is generated from the amino acids generated from the scoring function. Then using the amino acids from the probability distribution one generates a secondary library. This method can be used for all proteins.

There is no undue experimentation since the specification enables one skilled in the art to practice the invention using the specifically recited steps in the claims. The Examiner refers to Cys, Pro and Gly not being used in an Example in the specification. Applicants' respectfully refer the Examiner to page 17, lines 30-35, where the specification discloses the basis behind using, or not using certain amino acids in certain situations. To one skilled in the art of protein design, this is not undue experimentation but a

design choice. With respect to the Examiner's comments regarding SO_2 and water being removed, Applicants' respectfully refer the Examiner to page 15, lines 6-25 for the discussion on backbone structure preparation, as well as the discussion on backbone preparation above.

Applicants respectfully point to *In re Goffe*, 191 USPQ429 (CCPA 1976), where the court stated:

> "For all practical purposes, the Board would limit Appellant to claims involving the specific materials disclosed in the examples, so that a competitor seeking to avoid infringing the claims would merely have to follow the disclosure in the subsequently issued patent to find a substitute. However, to provide effective incentives, claims must adequately protect inventors. To demand that the first to disclose shall limit his claims to what he has found to work or to materials which meet the guidelines specified for "preferred" materials in a process such as the one herein involved would not serve the constitutional propose of promoting progress in the useful arts."

Additionally, in *In re Angstadt*, 190 USPQ 214, 218 (CCPA 1976), the court further stated:

> "Appellants have apparently not disclosed every catalyst which will work; they have apparently not disclosed every catalyst which will not work. The question, then, is whether in an unpredictable art, section 112 requires disclosure of a test with every species covered by a claim. To require such a complete disclosure would apparently necessitate a patent application or applications with "thousands" of examples or the disclosure of "thousands" of catalysts along with information as to whether each exhibits catalytic behavior resulting in the production of hydroperoxides. More importantly, such a requirement would force an inventor seeking adequate patent protection to carry out a prohibitive number of actual experiments. This would tend to discourage inventors from filing patent applications in an unpredictable area since the patent claims would have to be limited to those embodiments which are expressly disclosed."

Therefore, in conclusion, Applicants submit that the Specification taken in conjunction with the state of the art at the time the invention was filed fully enables a person skilled in the art to practice the method of the invention without undue experimentation. Applicants respectfully request reconsideration and withdrawal of the rejection.

Rejection under 35 USC §112, second paragraph

Claims 1-6 are rejected under 35 USC §112, second paragraph as being indefinite for failing to particularly point out and distinctly claim the subject matter, which the applicant regards as the invention.

A). The Examiner rejects claim 1 as being incomplete for omitting essential steps. Claim 1 has been amended to clarify the generation of the primary library. The use of a force field calculation produces a probability distribution table is also clarified. Further, one skilled in the art would understand how this works in view of the specification. The probability distribution table is used to generate a

secondary library of sequences, rather than a secondary sequence, as stated by the Examiner at Page 7, penultimate line. The force field calculation is applied to generate the primary library and a resulting probability distribution table of the variants, which in the simplest embodiments entails a simple counting process of the frequency of occurrence of each amino acid at each primary variant position in a list of low energy sequences (see Specification, page 58, lines 20-22 and table 3).

Examiner's use of the term "probability" is taken out of context. The term is "probability distribution table" or "probability distribution" and both are terms of art. This technique may be used to define a table (see Table 3 in the examples as one example) of the probability distribution of a set of amino acids at a particular position of a target protein. The use of a "probability distribution table" is not ambiguous since it is a defined table that is well defined in the specification and well known in the prior art by those skilled in the art. The term "probability" is not used in the "uncertainty" context as stated by Examiner. Moreover, the term is not used alone but in combination with the expression well known in the art "probability distribution" or "probability distribution table".

As stated in the MPEP §2173.05(a):

> The meaning of every term used in a claim should be apparent from the prior art or from the specification and drawings at the time the application is filed. Applicants need not confine themselves to the terminology used in the prior art, but are required to make clear and precise the terms that are used to define the invention whereby the metes and bounds of the claimed invention can be ascertained. During patent examination, the pending claims must be given the broadest reasonable interpretation consistent with the specification. *In re Morris*, 127 F.3d 1048, 1054, 44 USPQ2d 1023, 1027 (Fed. Cir. 1997); *In re Prater*, 415 F.2d 1393, 162 USPQ 541 (CCPA 1969). See also MPEP § 2111 - § 2111.01. When the specification states the meaning that a term in the claim is intended to have, the claim is examined using that meaning, in order to achieve a complete exploration of the applicant's invention and its relation to the prior art. *In re Zletz*, 893 F.2d 319, 13 USPQ2d 1320 (Fed. Cir. 1989).

In reviewing a claim for compliance with 35 U.S.C. §112, the Examiner must consider the claim as a whole to determine whether the claim apprises one of ordinary skill in the art of its scope and, therefore, serves the notice function required (See MPEP §2173.02). If the claims, read in light of the specification, reasonably apprise those skilled in the art both of the utilization and scope of the invention, and if the language is precise as the subject matter permits, the statute demands no more. [emphasis added].

Applicants have corrected the terms "scaffold protein sequences" and "secondary sequences" have been modified to make these terms more consistent in claim 1.

B). Claim 2 has been modified to clarify the Examiner's regarding the synthesizing steps. C). Claim 3 has been modified to clarify Examiner's concerns. D). Claim 5 has been amended to clarify the term "correspond". E). Claim 6 has been canceled making this rejection moot.

In light of the foregoing arguments, Applicants respectfully request the reconsideration and withdrawal of the rejection of Claims 1-6.

Double Patenting

Claims 1-6 have been provisionally rejected under the judicially created doctrine of obviousness type double patenting as being unpatentable over claims 1-9 of copending application No. 09/782,004. Applicants respectfully point out that the instant application is a divisional filing of the '004 case. The Examiner found that the claims in the '004 application were patentably distinct from the claims in the instant application. In response to the Examiner's parenthetical statement requesting that Applicant set a demarcation line among the numerous copending applications, it is respectfully submitted that many of the copending cases were the subject of the Examiner's restriction requirements and that the demarcation was established by the Examiner. If the Examiner no longer believes that the claims in these cases are patentably distinct, Applicants' respectfully request that the restriction requirement be withdrawn and the claims from all of the copending cases combined and considered together. Applicants would be happy to submit these claims with a linking genus claim. It is respectfully requested that this provisional rejection be withdrawn.

Claims 1-6 have been provisionally rejected under the judicially created doctrine of obviousness type double patenting as being unpatentable over claims 1-8 of US Patent No. 6,403,312. Applicant has enclosed a TERMINAL DISCLAIMER TO OBVIATE A DOUBLE PATENTING REJECTION in response to this rejection.

Claim Rejections – 35 USC § 103(a)

Claims 1-6 have been rejected under 35 USC §103 (a) as being unpatentable over Mayo (WO98/47089 is the foreign counterpart to US Patent No. 6,188,965 issued February 13, 2001) in view of Applicants' disclosure of known prior art.

To establish a prima facie case of obviousness, three basic criteria must be met: 1) suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art to modify or combine reference teachings; 2) there must be a reasonable expectation of success; and 3) the prior art reference must teach or suggest all the claim limitations. (See MPEP §2142).

The Office Action states "Mayo discloses e.g., page 15, lines 4-7; page 46, line 34 up to page 47, line 10 a method of creating a secondary sequence library with the side chains described as rotamers using force field calculation in generating a secondary structure for protein variants."

Applicants respectfully disagree with Examiner's assessment of this reference. In the section identified by the Examiner, Mayo is describing a validation of a scoring function, in this case, a van der Waals and solvation function. A secondary library is not generated, let alone a primary library, which is the methodology specifically defined in the claims.

Mayo used a force field calculation to generate energies of 75 previously defined and pre-existing full-length proteins. The proteins used were not "designed" in this case, nor was a primary library actually generated. The energies were generated so as to validate the use of a force field in a computational design program. Thus, while Mayo discloses inputting coordinates into a system, no library is generated from use of a force field, as energies were merely identified.

As further clarification, Applicant's respectfully point out that a rotamer library is not a primary or a secondary library. A rotamer library is a set of conformers that may be applied to a target protein to generate a primary library or secondary library. However, the term "library" with rotamer is not a "library" used in the sense of the instant claims.

In the above example, no rotamer libraries were used to generate a diverse library of primary sequences. The sequences were already identified; Mayo input the coordinates of these pre-existing molecules and then generated the various energies of each sequence using a van der Waals force field as a means of validating this scoring parameter.

While the Mayo reference, taken in its entirety, discloses the use of a rotamer library to design modified protein sequences, it does not disclose, teach or suggest, a secondary library, nor does it disclose, teach or suggest probability distribution nor probability distribution tables. Thus, Mayo does not disclose a secondary library as specifically recited in Applicants' claims, nor a probability distribution, also specifically recited in Applicants' claims.

The Examiner also recites "Applicants' disclosure of known prior art" as part of the 103(a) rejection. One of ordinary skill would understand the cited "DNA shuffling" reference and other well-known mutagenesis techniques, teach away from the use of a rational computational design algorithm, as used in the instant application. DNA shuffling is a random technique to generate libraries. There is no rational computational design used to generate a library. Thus, "Applicants' disclosure of known prior

art" does not teach the expressly recited steps of generating a secondary library, nor does this disclosure suggest or teach use of computational design.

With respect to the first criterion for a prima facie case of obviousness, there is no teaching, suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art to modify or combine reference teachings. The Examiner acknowledges Mayo does not disclose the synthesis of the protein or the nucleotides that would encode the protein. However, since Mayo does not generate a secondary library either, the combination of these teachings does not make obvious to one skilled in the art, Applicant's claimed invention.

Furthermore, there is no suggestion or motivation to modify or combine the teachings with Mayo, since the DNA shuffling technique is a random experimental technique that generates a library of unknown sequences that are undefined until these sequences are sequenced. There is no suggestion to combine these teachings since one is a random technique and Applicant's method is a rational one. Further to combine the references would destroy the prior art teachings of each. Therefore, the first prong of the analysis has not been met.

The second criterion, a reasonable expectation of success, has been demonstrated in the several working examples included in the application as filed. See Specification at pages 57-64. Additional support for the expectation of success may be found in the publications and patents recited in the response to the rejection under 35 USC § 101.

Finally, the prior art reference must teach or suggest all the claim limitations. As discussed above, neither prior art reference teaches or suggests all the claim limitations of the present invention. As discussed above, Mayo does not disclose generating a library of protein variants, and certainly not a secondary library. With respect to the prior art cited by the Examiner, e.g. "DNA shuffling," the reference does not disclose generating a probability distribution table of amino acid residues in a plurality of variant positions utilizing a force field, as required by the present invention. In light of the above-facts, neither of the cited prior art references teaches or suggests all the claim limitations and do not support the third criterion.

"A prior art reference must be considered in its entirety, i.e., as a whole, including portions that would lead away from the claimed invention. W.L. Gore & Associates, Inc. v. Garlock, Inc., 721 F.2d 1540, 220 USPQ 303 (Fed. Cir. 1983), cert denied, 469 U.S. 851 (1984)." As stated above, the combination of a rational approach and a random approach would destroy the teachings of each reference. The references teach away from each other and also the combination of them together.

Applicants respectfully submit, in light of the foregoing discussion, neither reference supports a finding that a prima facie case of obviousness has been established against the present invention.

Claims 1-6 have been rejected under §103(a) as being unpatentable over Mayo (WO 09/47089) in view of applicants' disclosure of known prior art. The Office Action states "Mayo discloses e.g., page 15, lines 4-7; page 46, line 34 up to page 47, line 10 a method of creating a secondary sequence library with the side chains described as rotamers using force field calculation in generating a secondary structure for protein variants."

Applicants' respectfully disagree because as discussed above, the cited reference neither suggests or teaches the generation of libraries per se, and does not teach the generation of secondary libraries.

Furthermore, the present invention may be distinguished from the cited reference because there is no suggestion or teaching of generating a secondary library from secondary sequences, differing from the primary sequence(s). Therefore, the claims of the present invention are not anticipated or made obvious by the cited reference because each and every element as set forth in the claim is not found, either expressly or inherently described, in a single prior art reference. Nor are the elements not set forth in the reference, suggested in the reference. There is no suggestion of combination of "gene shuffling" with Mayo by Mayo. In light of the foregoing, Applicants respectfully request reconsideration and withdrawal of the claim rejections.

The Applicants submit that in light of the above-amendment and argument, the claims are now in condition for allowance and an early notification of such is respectfully solicited.

The Examiner is invited to contact the undersigned at (415) 781-1989 if any issues may be resolved in that manner.

Respectfully submitted,
~~DORSEY~~ & WHITNEY LLP

Dated: September 24, 2004

By: Renee M Kosslak

Four Embarcadero Center
Suite 3400
San Francisco, California 94111-4187
Telephone: (415) 781-1989
Fax No. (415) 398-3249

Renee M. Kosslak, Reg. No. 47,717
for Robin M. Silva, Reg. No. 38,304

Combinatorial protein design

Jeffery G Saven

Combinatorial protein libraries permit the examination of a wide range of sequences. Such methods are being used for *de novo* design and to investigate the determinants of protein folding. The exponentially large number of possible sequences, however, necessitates restrictions on the diversity of sequences in a combinatorial library. Recently, progress has been made in developing theoretical tools to bias and characterize the ensemble of sequences that fold into a given structure – tools that can be applied to the design and interpretation of combinatorial experiments.

Addresses
Department of Chemistry, University of Pennsylvania, 231 South 34 Street, Philadelphia, Pennsylvania 19104, USA;
e-mail: saven@sas.upenn.edu

Current Opinion in Structural Biology 2002, **12**:453–458

0959-440X/02/$ – see front matter

Introduction

The discovery and design of novel proteins can lead to new, potentially practical proteins and can also enhance our understanding of protein biochemistry. Designing well-structured, soluble proteins is difficult, however, because of their complexity. Such proteins are large (tens to hundreds of amino acid residues) and have many variables that specify the folded state, including sequence, backbone topology and sidechain conformation. Design involves identifying those sequences that fold into a given structure from a huge ensemble of possible sequences. This search is aided, in part, by the large degree of consistency seen in folded proteins. On average, a folded structure is well packed, hydrophobic residues are sequestered from solvent and most potential hydrogen bond interactions are satisfied. This consistency, however, is often complex, may have little simplifying symmetry and involves predominantly noncovalent interactions. Such interactions are some of the most difficult to accurately quantify. As such, estimating the free energies associated with mutation or structural ordering remains a subtle area of computational research. Nonetheless, many molecular potentials do contain a 'best parameterization' of many of the interatomic interactions and forces that we know are important for stabilizing proteins. In some cases, such potentials have been used with striking success in protein design [1••]. Given that these potentials are necessarily approximate, however, one promising approach is to use the partial information contained in these functions in a probabilistic manner. A probabilistic or statistical approach is also appropriate for characterizing the full variability of sequences that fold to a common structure, because there are likely to be an enormous number of such sequences. Such statistical methods can be applied in 'shotgun' approaches to *de novo* protein design. Combinatorial experiments create and assay many sequences in order to overcome shortcomings in our understanding of folding or other molecular properties. Even though combinatorial methods can address large numbers of sequences (10^4–10^{12}), these numbers are still infinitesimal in comparison to the numbers of possible sequences (e.g. $20^{100} \approx 10^{130}$ for a 100-residue protein). Thus, methods for winnowing and focusing sequence space are a vital component of combinatorial protein design. Herein, I briefly discuss combinatorial methods for full sequence design. I also review recent theoretical developments in characterizing sequence ensembles — developments that can be applied to the design and interpretation of combinatorial experiments.

Directed protein design

There has been much effort — and success — in developing computational methods for 'directed' protein design. By 'directed protein design', I mean the identification of a sequence (or a small set of sequences) that is likely to fold into a predetermined backbone structure. Each such sequence can then be synthesized to confirm its folded structure and other molecular properties. Early efforts in design identified proteins with substantial order, but not necessarily well-defined tertiary structure [2]. Because an enormous number of sequences are possible even for small proteins (<50 residues), computational methods have dramatically accelerated successful design. Typically, such methods are implemented as an optimization process, whereby amino acid identity and sidechain conformation are varied in order to optimize a scoring function that quantifies sequence/structure compatibility. Exhaustive searching of all m^N possible sequences (where m is the number of different amino acid types or 'states' per residue and N is the number of residues in a target protein structure) is feasible only if a small number of residues N are allowed to vary or if the number of amino acids m is greatly reduced. If, in the optimization process, the different sidechain conformations (rotamer states) of each amino acid are also considered (see [3]), the complexity of the search increases still further, because m, the number of possible 'states' per residue, increases by a factor of ten or more. Although complete enumeration is typically not feasible, sequence space can be sampled in a directed manner in order to find optimal (or nearly optimal) sequences. Stochastic methods, such as genetic algorithms or simulated annealing, involve searching sequence space in a partially random fashion; on average, the search progressively moves toward better scoring (lower energy) sequences [4,5]. The partially random nature of the search permits escape from local minima in the sequence/rotamer landscape. Using a simplified model, the Takada and Tamura groups have included information about unfolded structures (negative design) in a stochastic search for a sequence with a 'funneled conformational energy landscape' [6]. One

47-residue three-helix bundle protein so selected has CD and NMR spectral features of folded proteins (W Jin, O Kambara, H Sasakawa, A Tamura, S Takada, personal communication). When applied to atomically detailed representations, the stochastic methods focus primarily on repacking the interior of a structure with hydrophobic residues [7] and have been applied to the wild-type structures of 434 Cro [8], ubiquitin [9], the B1 domain of protein G [10•], the WW domain [1••] and helical bundles [11,12]. Although, in many cases, these methods have identified experimentally viable sequences [1••,13], stochastic search methods need not identify global optima [14•]. For potentials comprising only site and pair interactions, elimination methods such as 'dead end elimination' can find the global optimum [14•,15–17]. Such methods successively remove individual amino acid rotamer states that cannot be part of the global optimum until no further states can be eliminated. The Mayo group applied such methods to automate the full sequence design of both a 28-residue zinc finger mimic [18] and, after predetermining hydrophobic and polar sites, a 51-residue homeodomain motif [19•]. The group has also redesigned portions of a variety of proteins [20–22]. Functional properties such as metal binding or catalysis may also be included as elements of the design process [23,24•]. The elements and algorithms of directed protein design have been the subject of several recent reviews [1••,25,26•].

Despite some striking successes, computational methods for directed design have limitations with respect to both identifying folding sequences and characterizing the features of protein sequences that share a common structure. Stochastic methods, such as simulated annealing or genetic algorithms, can be applied to large proteins and permit many sites to be varied simultaneously, but the computational times and resources required for such calculations are extensive, even for small proteins. When used as optimization methods, directed approaches will necessarily be sensitive to the energy or scoring function used. All energy functions in use in protein design, however, are necessarily approximate and uncertainties in the energy function may not merit the search for global optima. Furthermore, many naturally occurring proteins are not optimized. In fact, most proteins are only marginally stable (e.g. ΔG^{o} <10 kcal/mol for folding) [27]. In addition, sequences that function, for example, those that bind another molecule, need not be the global optimum with respect to structural stability. Although stochastic methods can sample such suboptimal sequences, in general an exponentially large number of them will be possible and such sampling will be time consuming. Thus, it is important to develop methods complementary to those used for directed protein design — methods that reveal the features of sequences that are likely to fold into a particular structure but that may not be structurally 'optimal'. Such computational methods will have application to a new class of protein design studies, combinatorial experiments, in which large numbers of proteins may be simultaneously synthesized and screened.

Combinatorial design

Combinatorial design provides a complementary approach to directed design for understanding sequence/ structure compatibility and discovering novel sequences that fold into a specific structure. Combinatorial methods are powerful tools for cases in which we have an incomplete understanding of molecular properties. In protein combinatorial design experiments, large numbers of sequences (libraries) are screened for evidence of folding into a predetermined structure. A combinatorial experiment has two key elements: creating a library with a desired degree of diversity and assaying for sequences with 'protein-like' properties in terms of their structure or function. Depending upon how the diversity is generated and assayed, experiments of this type can explore a large number of sequences, up to 10^{12} [28•]. Certainly, such methods can be used to discover 'hits', that is, a few sequences that are especially stable or that are unusually strong in their function or binding properties. In addition, combinatorial experiments readily generate a sequence ensemble. Thus, using combinatorial experiments, we can potentially 'expand the protein sequence database' and the diversity of these additional sequences will be at the control of the researcher. Features important to folding (and other properties) may be explored in a way that is decoupled from the evolutionary requirements of nature's proteins. For example, these methods have been used to identify helical proteins [29–31], ubiquitin variants [32], self-assembled protein monolayers [33], proteins with amyloid-like properties [33], metal-binding peptides [34] and stable interhelical oligomers [35]. Several excellent reviews of combinatorial experiments have appeared recently [36,37,38•,39••].

The complexity of combinatorial experiments implies that limitations must be placed on the sequences, because the number that can be created and screened (10^{6}–10^{12}) is infinitesimal compared to the number possible (e.g. 10^{130}). Limitations on sequence properties are often guided by qualitative chemical considerations, but quantitative computational methods will be helpful in designing and interpreting combinatorial experiments.

The Hecht group has probed the extent to which the patterning of hydrophobic and hydrophilic residues can successfully reduce complexity in combinatorial design. While maintaining the periodicity of α helices and β sheets in particular tertiary structures, such patterning is applied in order to expose hydrophilic residues to solvent and to sequester hydrophobic residues in the interior of the protein. Early targets were helical proteins; a fiducial 74-residue four-helix bundle was the template structure [40]. Such a structure has more than $20^{74} \approx 10^{96}$ possible sequences. After binary patterning, five hydrophobic and six hydrophilic amino acids were permitted at 24 interior and 36 exterior positions, respectively, thus reducing the total number of possible sequences to 10^{41}. From a protein library consistent with this binary patterning, a set of 50 correctly expressed sequences was selected for further

study. Around half of the 50 sequences isolated are protein-like in many respects [30], including their thermal denaturation [41]. About half the isolated sequences also bind heme [29] and many of these display carbon monoxide binding [42•] or peroxidase activity [43]. This is surprising given that such functions were not part of the design or selection of the sequences. In a second-generation design, the group added six residues to each of the four helices of one of the most protein-like sequences. The additional residues were combinatorially patterned, as in the original experiment [39••]. For these 102-residue sequences, the free energies of folding are increased 2–3-fold and the NMR data suggest well-determined structures. Using binary patterning of hydrophobicity consistent with an amphilic β sheet [44], the Hecht group has also identified proteins that aggregate to form amyloid fibrils [45] and crafted monomeric β proteins by introducing a nonpolar lysine mutation at the 'edge' strand of the target β sheet [46••].

Despite the striking results from hydrophobic patterning, more detailed methods for library design are merited. Many of the hydrophobically patterned sequences that appear well structured are not sufficiently soluble for NMR structure determination [46••] and, as a result, little is known concerning their structures at the atomic scale. Not all of the α-helical sequences exhibit the sharp thermal transition seen in natural proteins (usually associated with a large ΔH of folding). Such sequences may not possess well-packed interiors [41]. In natural proteins, the side-chains of most interior residues are well determined, as opposed to the variability that is obtained using hydrophobic patterning alone and that is observed in many *de novo* designed proteins [13,18]. A more fine-grained dictation of the amino acid identities is probably necessary for obtaining libraries that are rich in sequences with well-defined structures. Moreover, a more detailed specification of amino acid identities yields fewer sequences than hydrophobic patterning alone and further reduces the complexity of the library.

Theories of combinatorial libraries

Surveying the complete sequence landscape of proteins seems, at first glance, intractable to both experiment and computation. In addition to the enormous number of possible sequences, many examples exist in nature of dissimilar sequences folding to essentially the same structure. Hence, sequence properties are nontrivial and proteins sharing a common structure can be nonlocal in sequence space. Nonetheless, computational methods permit us to estimate the properties, particularly the amino acid probabilities, of sequences consistent with a target structure.

Repeated use of directed search methods can estimate the properties of an ensemble of sequences. Desjarlais and co-workers have used independent runs of their sequence prediction algorithm across an ensemble of closely related structures all consistent with a particular fold (JR Desjarlais *et al.*, personal communication). For each structure, an optimal 'nucleating' sequence is identified and subsequently the sequence/rotamer variability is explored throughout the structure. The method identifies effective reduced partition sums for each sequence/rotamer state and amino acid probabilities may be obtained at each residue position. The number of sequences decreases with stability, so the degree of complexity can be tuned by varying a cutoff in the effective free energies of the sequences. The method has been used to identify sequences consistent with the fold of a WW domain, a small β-sheet protein [1••], some of which are currently being experimentally characterized.

The amino acid frequencies can also be determined directly, using a statistical theory of combinatorial libraries [47,48••,49••]. Ideas from statistical mechanics are used to address the number and composition of sequences that are consistent with a particular backbone structure. The theory addresses the whole space of available compositions, not just the small fraction that is accessible to experiment and to computational enumeration and sampling. The theory takes as input a target backbone structure and a scoring or energy function for quantifying sequence/structure compatibility. Global and local features can be prespecified using constraints on the sequences. For example, such constraints can be used to determine the energy the sequences assume in the target structure, the patterning of amino acids and the number of each amino acid present (composition). The theory yields estimates of both the number of sequences consistent with these constraints and the amino acid probabilities at each residue position. These residue-specific probabilities are the most probable such set and are determined — as in statistical mechanics — by maximizing an effective entropy, whereby this maximization is subject to constraints. Just as in thermodynamics, the judicious use of constraints can be used to reduce the entropy or the number of possible sequences. Thus, these methods provide a systematic means to focus the library, winnowing numbers such as 10^{130} to numbers that are experimentally manageable, for example, 10^6. The theory agrees well with exact results obtained with lattice models of proteins [47,48••]. This method has been extended to realistic representations of proteins, in which the effects of sidechain packing are included in an atom-based manner [49••]. The calculated sequence probabilities of the immunoglobulin light chain binding domain of protein L are in agreement with the frequencies observed in combinatorial phage display experiments [50,51]. These statistical methods have several advantages. They may be applied to much larger proteins (N>100 residues) and permit much larger sequence variation than many directed methods. They are sufficiently rapid that many backbone structures may be considered and those features that are robust with respect to minor structure modifications may be identified. Importantly, such methods provide perhaps the most natural input for a combinatorial experiment, the probabilities of the amino acids at each position among the sequences of a library. These amino acid

probabilities can also be used to identify specific amino acid sequences, which can then be synthesized; a consensus sequence comprising the most probable amino acid at each site can be selected or the probabilities can be used to bias a stochastic search for viable sequences (J Zou, JG Saven, unpublished data).

If the energy of the target state is one of the constraints, the statistical method reduces to an effective mean field theory. Mean field theories have seen extensive application in physical science and in biomolecular theory [52], and to protein evolution and natural sequence variability ([53]; H Kono, JG Saven, unpublished data). Voigt *et al.* [14•] have compared mean field theories with directed search methods for identifying ground state sequence/rotamer combinations in protein design. They found that, although often more rapid, mean field theories do not always identify such ground states. Interestingly, Voigt *et al.* applied the mean field theory to large proteins (subtilisin E and T4 lysozyme) to determine local site entropies, s_i, where $\exp(s_i)$ quantifies the effective number of amino acids allowed at residue i in a structure [54••,55]. Sites with large values of s_i, those most tolerant to mutation [56], are likely to support substitutions that improve stability or function when *in vitro* evolution experiments are used to explore sequence space [37]. For such experiments, the mutation rate is low enough that multiple mutations of strongly interacting sites are rare. Thus, mutations that improve 'fitness' are most likely to accumulate at sites that are the most 'decoupled' from other sites. Such mutations can potentially be targeted for variation in an *in vitro* evolution experiment.

Conclusions

Much recent progress has been seen in the design and discovery of new proteins, and combinatorial approaches are accelerating the pace. Such methods are most useful when our quantitative understanding of important protein properties, such as stability and catalytic activity, is limited. Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but, with combinatorial experiments, an enormous diversity of sequences can be generated at the control of the researcher. Detailed questions can be addressed, such as the utility of hydrophobic patterning or of predetermining particular sites for amino acid variation. Theory and simulation will continue to aid the design and interpretation of combinatorial experiments. Such methods will also facilitate the exploration of what is possible with the amino acids: how diverse is the set of all possible sequences that fold to a particular structure and what structures not yet seen in nature can be crafted with the amino acids? Such methods will perhaps have an even more profound impact on designing nonbiological foldamers [57••], structures about which we have much less empirical information than we do about biopolymers.

Acknowledgements

The author acknowledges support from the National Science Foundation (CHE 98-16497 and CHE 99-84752). JGS is a Cottrell Scholar of Research Corporation and is an Arnold and Mabel Beckman Foundation Young Investigator.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Kraemer-Pecore CM, Wollacott AM, Desjarlais JR: **Computational protein design.** *Curr Opin Chem Biol* 2001, **5**:690-695.
This is a compact but excellent review on recent progress in computational methods for protein design. The authors also discuss recent efforts in designing the WW domain, a small β protein.

2. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF: **Protein design: a hierarchic approach.** *Science* 1995, **270**:935-941.

3. Dunbrack R: **Rotamer libraries.** *Curr Opin Struct Biol* 2002, **12**:in press.

4. Shakhnovich EI, Gutin AM: **A new approach to the design of stable proteins.** *Protein Eng* 1993, **6**:793-800.

5. Jones DT: ***De novo* protein design using pairwise potentials and a genetic algorithm.** *Protein Sci* 1994, **3**:567-574.

6. Onuchic JN, Luthey-Schulten Z, Wolynes PG: **Theory of protein folding: the energy landscape perspective.** *Annu Rev Phys Chem* 1997, **48**:545-600.

7. Hellinga HW, Richards FM: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, **91**:5803-5807.

8. Desjarlais JR, Handel TM: ***De-novo* design of the hydrophobic cores of proteins.** *Protein Sci* 1995, **4**:2006-2018.

9. Johnson EC, Lazar GA, Desjarlais JR, Handel TM: **Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin.** *Structure* 1999, **7**:967-976.

10. • Jiang X, Farid H, Pistor E, Farid RS: **A new approach to the design of uniquely folded thermally stable proteins.** *Protein Sci* 2000, **9**:403-416.
The authors use a novel scoring function for the design of hydrophobic interiors. In addition to steric interactions, the function includes parameterizations of changes in the heat capacity and the conformational entropy upon folding. Simulated annealing was used to optimize the score. The backbone and exterior residue identities were constrained. In tests on two small proteins, in which 10–11 interior residues were varied, the native sequence was regenerated, as well as the sequences of known stable variants. Interestingly, previously designed sequences with low stability and weak cooperativity were not identified. In larger proteins tested, in which 32 and 63 residues were varied, sequence/rotamer combinations close to native were identified.

11. Jiang X, Bishop EJ, Farid RS: **A *de novo* designed protein with properties that characterize natural hyperthermophilic proteins.** *J Am Chem Soc* 1997, **119**:838-839.

12. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF: **From coiled coils to small globular proteins: design of a native-like three-helix bundle.** *Protein Sci* 1998, **7**:1404-1414.

13. Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF: **Solution structure and dynamics of a *de novo* designed three-helix bundle protein.** *Proc Natl Acad Sci USA* 1999, **96**:5486-5491.

14. • Voigt CA, Gordon DB, Mayo SL: **Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design.** *J Mol Biol* 2000, **299**:789-803.
The authors carefully compared different methods for sequence design, including simulated annealing, genetic algorithms, mean field methods and dead end elimination (DEE). DEE most reliably finds global minima, but the authors also note that the method may be limited to 30 amino acid sites for which full amino acid variability is permitted. The authors extrapolate the results to regimes to which DEE cannot be applied. They find that both mean field and annealing approaches perform best with core residues and less reliably with residues that are fully or partially solvent exposed.

15. Gordon DB, Mayo SL: **Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem.** *J Comput Chem* 1998, **19**:1505-1514.

16. Gordon DB, Mayo SL: **Branch-and-terminate: a combinatorial optimization algorithm for protein design.** *Structure* 1999, **7**:1089-1098.

17. Pierce NA, Spriet JA, Desmet J, Mayo SL: **Conformational splitting: a more powerful criterion for dead-end elimination.** *J Comput Chem* 2000, **21**:999-1009.

18. Dahiyat BI, Mayo SL: ***De novo* protein design: fully automated sequence selection.** *Science* 1997, **278**:82-87.

19. • Marshall SA, Mayo SL: **Achieving stability and conformational specificity in designed proteins via binary patterning.** *J Mol Biol* 2001, **305**:619-631.
The authors combined binary patterning with atomistic sidechain interactions to identify folding sequences of a homeodomain. Preliminary calculations were done using 'generic amino acids' to identify those sites that are most appropriate for polar or hydrophobic residues. Subject to this binary patterning, a directed search was then performed using dead end elimination. Interestingly, the authors identified an optimal binary patterning, whereby adding or subtracting hydrophobic residues adversely affects folding to stable monomers.

20. Malakauskas SM, Mayo SL: **Design, structure, and stability of a hyperthermophilic protein variant.** *Nat Struct Biol* 1998, **5**:470-475.

21. Strop P, Mayo SL: **Rubredoxin variant folds without iron.** *J Am Chem Soc* 1999, **121**:2341-2345.

22. Shimaoka M, Shifman JM, Jing H, Takagi L, Mayo SL, Springer TA: **Computational design of an integrin I domain stabilized in the open high affinity conformation.** *Nat Struct Biol* 2000, **7**:674-678.

23. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A: ***De novo* design and structural characterization of proteins and metalloproteins.** *Annu Rev Biochem* 1999, **68**:779-819.

24. • Bolon DN, Mayo SL: **Enzyme-like proteins by computational design.** *Proc Natl Acad Sci USA* 2001, **98**:14274-14279.
The authors designed non-native enzymatic activity into a thioredoxin fold. The authors computationally identified promising active sites on the scaffold. The sequence was designed to stabilize the transition state of a hydrolysis reaction. The enzymes so designed had activity well above background.

25. Street AG, Mayo SL: **Computational protein design.** *Structure* 1999, **7**:R105-R109.

26. • Saven JG: **Designing protein energy landscapes.** *Chem Rev* 2001, **101**:3113-3130.
The author reviews recent progress in protein design from the perspective of the energy landscape theory of folding. In the context of theory, models and real systems, different issues involved in design are discussed, including target structures, energy functions, foldability criteria, search methods and the size of the amino acid alphabet.

27. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A: **ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0.** *Nucleic Acids Res* 2002, **30**:301-302.

28. • Keefe AD, Szostak JW: **Functional proteins from a random-sequence library.** *Nature* 2001, **410**:715-718.
A fascinating study on a random search for 'function' among random amino acid sequences. Using combinatorial methods the authors have pioneered, ATP-binding proteins were selected from a library of 10^{12} sequences.

29. Rojas NRL, Kamtekar S, Simons CT, Mclean JE, Vogel KM, Spiro TG, Farid RS, Hecht MH: ***De novo* heme proteins from designed combinatorial libraries.** *Protein Sci* 1997, **6**:2512-2524.

30. Roy S, Ratnaswamy G, Boice JA, Fairman R, McLendon G, Hecht MH: **A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties.** *J Am Chem Soc* 1997, **119**:5302-5306.

31. Roy S, Helmer KJ, Hecht MH: **Detecting native-like properties in combinatorial libraries of *de novo* proteins.** *Fold Des* 1997, **2**:89-92.

32. Finucane MD, Tuna M, Lees JH, Woolfson DN: **Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries.** *Biochemistry* 1999, **38**:11604-11612.

33. Xu GF, Wang WX, Groves JT, Hecht MH: **Self-assembled monolayers from a designed combinatorial library of *de novo* beta-sheet proteins.** *Proc Natl Acad Sci USA* 2001, **98**:3652-3657.

34. Case MA, McLendon GL: **A virtual library approach to investigate protein folding and internal packing.** *J Am Chem Soc* 2000, **122**:8089-8090.

35. Arndt KM, Pelletier JN, Muller KM, Alber T, Michnick SW, Pluckthun A: **A heterodimeric coiled-coil peptide pair selected *in vivo* from a designed library-versus-library ensemble.** *J Mol Biol* 2000, **295**:627-639.

36. Zhao HM, Arnold FH: **Combinatorial protein design: strategies for screening protein libraries.** *Curr Opin Struct Biol* 1997, **7**:480-485.

37. Giver L, Arnold FH: **Combinatorial protein design by *in vitro* recombination.** *Curr Opin Chem Biol* 1998, **2**:335-338.

38. • Hoess RH: **Protein design and phage display.** *Chem Rev* 2001, **101**:3205-3218.
A comprehensive review of a commonly used method to generate and display combinatorial libraries of proteins and peptides.

39. •• Moffet DA, Hecht MH: ***De novo* proteins from combinatorial libraries.** *Chem Rev* 2001, **101**:3191-3203.
A review of recent work on the *de novo* combinatorial design of proteins, focusing primarily on the pioneering work of the Hecht group.

40. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH: **Protein design by binary patterning of polar and nonpolar amino-acids.** *Science* 1993, **262**:1680-1685.

41. Roy S, Hecht MH: **Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids.** *Biochemistry* 2000, **39**:4603-4607.

42. • Moffet DA, Case MA, House JC, Vogel K, Williams RD, Spiro TG, McLendon GL, Hecht MH: **Carbon monoxide binding by *de novo* heme proteins derived from designed combinatorial libraries.** *J Am Chem Soc* 2001, **123**:2109-2115.
Heme-assisted binding of a diatomic ligand turns out to be easier to find than expected within a library of sequences patterned to form a four-helix bundle. Eight combinatorially selected heme-binding sequences bind carbon monoxide with an affinity similar to that of myoglobin. The binding properties of the proteins aren't nearly as diverse as those seen among natural heme proteins, but these *de novo* sequences serve as a useful 'reference'.

43. Moffet DA, Certain LK, Smith AJ, Kessel AJ, Beckwith KA, Hecht MH: **Peroxidase activity in heme proteins derived from a designed combinatorial library.** *J Am Chem Soc* 2000, **122**:7612-7613.

44. West MW, Beasley JR, Hecht MH: **Collections of *de novo* beta-sheet proteins designed by binary patterning of polar and nonpolar amino acids.** *Protein Eng* 1997, **10**:38-38.

45. West MW, Wang WX, Patterson J, Mancias JD, Beasley JR, Hecht MH: ***De novo* amyloid proteins from designed combinatorial libraries.** *Proc Natl Acad Sci USA* 1999, **96**:11211-11216.

46. •• Wang WX, Hecht MH: **Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins.** *Proc Natl Acad Sci USA* 2002, **99**:2760-2765.
Previously, the authors had used hydrophobic patterning consistent with β sheets that intermolecularly align 'edge on' and found these sequences did indeed form the amyloid fibrils that were expected. In this paper, they break up these edge-on interactions with a hydrophilic residue (lysine) at each edge of the β sheet. These sequences are indeed monomeric and appear to be well structured according to CD and NMR peak dispersion. These are the first examples of combinatorial β-protein design.

47. Saven JG, Wolynes PG: **Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules.** *J Phys Chem B* 1997, **101**:8375-8389.

48. •• Zou JM, Saven JG: **Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure.** *J Mol Biol* 2000, **296**:281-294.
The authors extend their statistical theory of sequence libraries to include negative design. The sequence space is resolved in multiple dimensions and the number of sequences is characterized according to the folded state energy and stability gap (the difference in energy between the folded state and an ensemble of unfolded conformations). Excellent agreement is observed between theoretical and exact lattice model results for both the numbers of sequences and the monomer probabilities.

49. •• Kono H, Saven JG: **Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure.** *J Mol Biol* 2001, **306**:607-628.
A statistical theory of combinatorial libraries developed for combinatorial experiments. The authors used an atom-based potential and rotamer states to identify the sequence probabilities consistent with a particular structure. An effective one-body energy was introduced that relates the hydrophobicity

or solvent-exposure propensity to local β-carbon density. The calculations give good results with regard to sidechain modeling. Calculations were done that are consistent with recent combinatorial experiments on protein L. Generally, the calculations are in good agreement with the observed amino acid frequencies, despite the sampling issues that are always a concern with these comparisons. (Only 20–40 sequences were sequenced in the experiments.)

50. Kim DE, Gu HD, Baker D: **The sequences of small proteins are not extensively optimized for rapid folding by natural selection.** *Proc Natl Acad Sci USA* 1998, **95**:4982-4986.

51. Gu H, Doshi N, Kim DE, Simons KT, Santiago JV, Nauli S, Baker D: **Robustness of protein folding kinetics to surface hydrophobic substitutions.** *Protein Sci* 1999, **8**:2734-2741.

52. Koehl P, Delarue M: **Mean-field minimization methods for biological macromolecules.** *Curr Opin Struct Biol* 1996, **6**:222-226.

53. Dokholyan NV, Shakhnovich EI: **Understanding hierarchical protein evolution from first principles.** *J Mol Biol* 2001, **312**:289-307.

54. •• Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computational method to reduce the search space for directed protein evolution.** *Proc Natl Acad Sci USA* 2001, **98**:3778-3783.
The authors used a mean field theory to determine each residue's structural tolerance to mutations. This tolerance is quantified by the residue's local sequence entropy, which is a measure of the effective number of amino acids that are structurally permitted at that site. For an *in vitro* directed evolution experiment, the authors suggest that mutations that enhance stability or activity are most likely to accumulate in these high entropy regions. Multiple compensating mutations are rare in such experiments, so mutations are most likely at sites that tolerate multiple amino acids. Calculations involving subtilisin E and T4 lysozyme are consistent with the mutations observed in directed evolution experiments.

55. Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computationally focusing the directed evolution of proteins.** *J Cell Biochem* 2001:58-63.

56. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**:56-68.

57. •• Hill DJ, Mio MJ, Prince RB, Hughes TS, Moore JS: **A field guide to foldamers.** *Chem Rev* 2001, **101**:3893-4011.
A comprehensive review of nonbiological folding molecules.

EXHIBIT F

BIOCHEMISTRY

DONALD VOET
University of Pennsylvania

JUDITH G. VOET
Swarthmore College

Illustrators:
IRVING GEIS
JOHN AND BETTE WOOLSEY
PATRICK LANE

WILEY
JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

To:

Our parents, who encouraged us,
Our teachers, who enabled us, and
Our children, who put up with us.

Cover Art: One of a series of color studies of horse heart cytochrome *c* designed to show the influence of amino acid side chains on the protein's three-dimensional folding pattern. We have selected this study to symbolize the discipline of biochemistry: Both are beautiful but still in process and hence have numerous "rough edges." Drawing by Irving Geis in collaboration with Richard E. Dickerson.

Cover and part opening illustrations copyrighted by Irving Geis.

Cover Designer: Madelyn Lesure

Photo Research: John Schultz, Eloise Marion

Photo Research Manager: Stella Kupferberg

Illustration Coordinator: Edward Starr

Copy Editor: Jeannette Stiefel

Production Manager: Lucille Buonocore

Senior Production Supervisor: Linda Muriello

Library of Congress Cataloging in Publication Data:

Voet, Donald.
Biochemistry / by Donald Voet and Judith G. Voet.
p. cm.
Includes bibliographical references.
ISBN 0-471-61769-5
1. Biochemistry. I. Voet, Judith G. II. Title.
QP514.2.V64 1990 89-16727
574.19'2—dc20 CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2

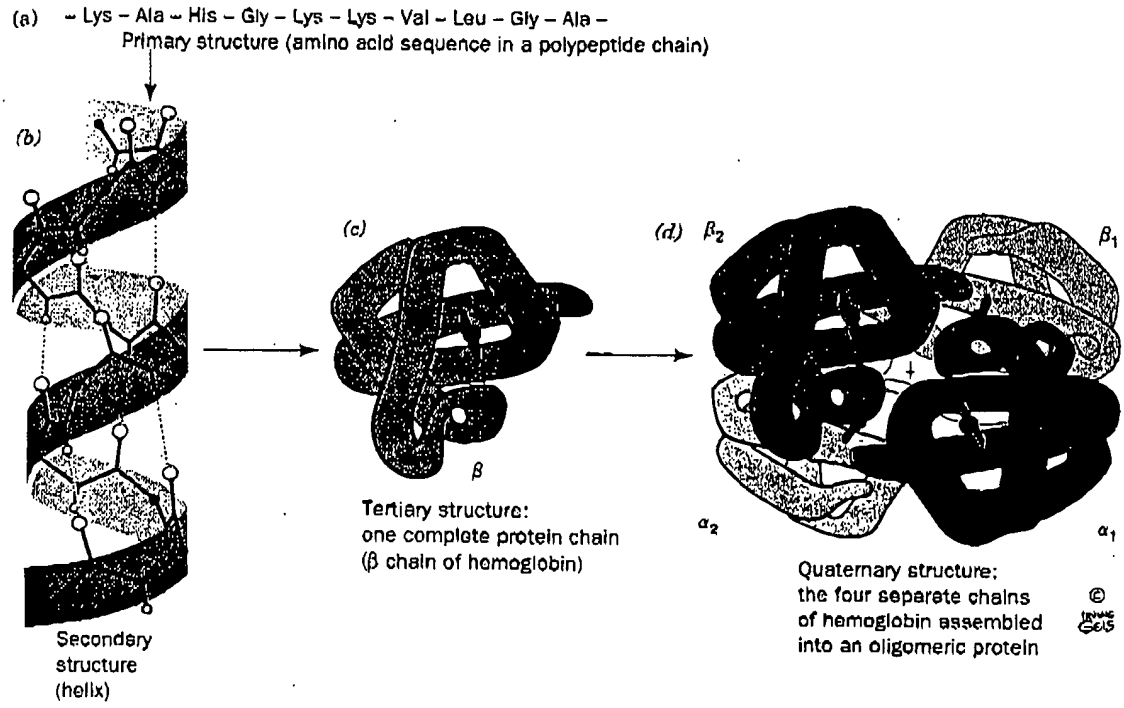

Figure 6-1
The structural hierarchy in proteins: (*a*) primary structure, (*b*) secondary structure, (*c*) tertiary structure, and (*d*) quaternary structure. [Figure copyrighted © by Irving Geis.]

Protein function can only be understood in terms of protein structure, that is, the three-dimensional relationships between a protein's component atoms. The structural descriptions of proteins, as well as those of other polymeric materials, have been traditionally described in terms of four levels of organization (Fig. 6-1):

1. A protein's **primary structure (1° structure)** is the amino acid sequence of its polypeptide chain(s).
2. **Secondary (2°) structure** is the local spatial arrangement of a polypeptide's backbone atoms without regard to the conformations of its side chains.
3. **Tertiary (3°) structure** refers to the three-dimensional structure of an entire polypeptide. The distinction between secondary and tertiary structures is, of necessity, somewhat vague; in practice, the term secondary structure alludes to easily characterized structural entities such as helices.
4. Many proteins are composed of two or more polypeptide chains, loosely referred to as **subunits**, which associate through noncovalent interactions and, in some cases, disulfide bonds. A protein's **quaternary (4°) structure** refers to the spatial arrangement of its subunits.

In this, the first of four chapters on protein structure, we discuss the 1° structures of proteins: How they are elucidated and their biological and evolutionary significance. We also survey methods of chemically synthesizing polypeptide chains. The 2°, 3°, and 4° structures of proteins which, as we shall see, are a consequence of their 1° structures, are treated in Chapter 7. In Chapter 8 we take up protein folding, dynamics, and structural evolution, and in Chapter 9 we analyze hemoglobin as a paradigm of protein structure and function.

1. PRIMARY STRUCTURE DETERMINATION

The first determination of the complete amino acid sequence of a protein, that of the bovine polypeptide hormone **insulin** by Frederick Sanger in 1953, was of enormous biochemical significance in that it definitively established that proteins have unique covalent structures. Since that time, the amino acid sequences of several thousand proteins have been elucidated. This extensive information has been of central importance in the formulation of modern concepts of biochemistry for several reasons: